Convergent evolution in primates and an insectivore

Dario Boffelli, Jan-Fang Cheng and Edward M. Rubin

Genome Sciences Department

Lawrence Berkeley National Laboratory

1 Cyclotron Rd, MS 84-171

Berkeley, CA 94720

Tel: 510-486-5072

Fax: 510-486-4229

e-mail: emrubin@lbl.gov

Abstract

The cardiovascular risk factor apolipoprotein(a) (apo(a)) has a puzzling distribution among mammals, its presence being limited to a subset of primates and a member of the insectivore lineage, the hedgehog. To explore the evolutionary history of apo(a), we performed extensive genomic sequence comparisons of multiple species with and without an apo(a) gene product, such as human, baboon, hedgehog, lemur and mouse. This analysis indicated that apo(a) arose independently in a subset of primates, including baboon and human, and an insectivore, the hedgehog, and was not simply lost by species lacking it. The similar structural domains shared by the hedgehog and primate apo(a) indicate that they were formed by a unique molecular mechanism involving the convergent evolution of paralogous genes in these distant species.

Introduction

Apo(a) is a protein found nearly exclusively in covalent association with low-density-lipoprotein (LDL) in the plasma of humans. A most unusual feature of apo(a) is its extremely different levels in humans, varying almost a thousand fold among individuals (between less than 0.1 mg/dl to more than 100 mg/dl). This remarkable concentration polymorphism is mostly associated with the apo(a) locus [1; 2] and is dependent on both the polymorphic structure of apo(a) [3] and on regulatory elements controlling this gene [4; 5].

The clinical relevance of apo(a) plasma levels relate to the direct association noted between it and the risk of developing coronary artery disease [6; 7; 8]. In spite of this well-established role in the pathogenesis of atherosclerosis, apo(a) has eluded attempts to define its essential biological functions. This is in part the result of the observation that no associated phenotypes have been detected in the large number of individuals having low to absent apo(a) plasma levels [9]. The structural similarity between apo(a) and its neighboring gene, plasminogen, suggests a possible mechanism by which apo(a) may promote the development of atherosclerosis through competition with plasminogen for fibrin binding [10].

An additional unusual feature of apo(a) is its peculiar and very limited distribution among mammals, being found only in old-world monkeys and hominoids, and also in a member of the insectivore family, the hedgehog [11]. Considering that the insectivore and primate lineages split at the beginning of mammal radiation [12; 13], the most likely evolutionary explanations for the appearance of the apo(a) gene in only a subset of primates and an insectivore involve the duplication of plasminogen followed by gene loss or convergent evolution. In a gene loss scenario, plasminogen would have duplicated before the split of

hedgehog ancestors from the other placental mammals and subsequently deleted in all species but those presently carrying the gene. For convergent evolution to explain apo(a) distribution among insectivores and primates, it would have arisen independently in each of these lineages through remodeling of plasminogen to acquire the functional properties of apo(a). A previous study, based on the partial hedgehog cDNA sequence analysis and reconstruction of the phylogenetic tree of plasminogen and apo(a), supported the hypothesis of separate duplication events, within the limitations of accurately calibrating and using the molecular clock to accurately estimate lineages branching times [14].

The analysis of the sequence of the genomic locus containing apo(a) and its flanking genes in several species and of the corresponding orthologous intervals in species without apo(a) offers the most direct approach to decipher the evolutionary relationship between primate and insectivore apo(a). While BAC libraries were available for primates with apo(a) and non-primate mammals lacking apo(a), in order to access this sequence in an insectivore and a primate lacking apo(a), we generated BAC libraries from genomic DNA of both the hedgehog and the prosimian lemur. Through the isolation and comparison of orthologous BAC clone sequences from hedgehog, mouse, lemur, baboon and human, we explored the evolutionary relationships of the apo(a)/plasminogen locus among these various species.

RESULTS AND DISCUSSION

Sequence analysis of the insectivore and primate apo(a)/plasminogen locus. The human apo(a) gene arose from the duplication of plasminogen, which is located within 40kb of apo(a) on chromosome 6 [15] (Fig. 2). Plasminogen contains a protease domain and five distinct structural domains of the kringle family (type I-V) while apo(a) contains two of plasminogen kringles (IV and V) and lacks proteolytic activity (Fig. 1). Other features of primate apo(a) distinguishing it from plasminogen include an unpaired cysteine on the last kringle IV that allows it to be covalently bound to apoB, the major protein component of the plasma lipoprotein LDL. In addition, while plasminogen has but a single copy of kringle IV, apo(a) has 8-40 copies of this domain. A basic attribute of the single kringle IV in plasminogen and the multiple copies of kringle IV in primate apo(a) is that they possess fibrin binding pockets conferring fibrin binding activity.

Similar to human, analysis of the BAC containing hedgehog apo(a) revealed its localization within 40 kb of plasminogen. The extensive sequence similarities between these two genes indicate that also hedgehog apo(a) likely arose by duplication of plasminogen. Examination of the predicted hedgehog plasminogen amino acid sequence showed it to be 86% similar to its human ortholog, resulting in highly homologous protein secondary structures containing a proteolytic domain and five distinct kringle domains (Fig. 1A). Comparing the human and hedgehog apo(a) genes to their respective plasminogen paralogs demonstrates informative similarities and differences (Fig. 1). Both versions of apo(a) appear to have been the product of the remodeling of their respective plasminogen genes resulting in the acquisition of an apoB-binding cysteine and of multiple copies of a single plasminogen kringle unit. In

addition, the protease activity of plasminogen is absent in both the insectivore and primate apo(a).

Important differences between human and hedgehog apo(a) include the specific plasminogen kringle amplified in apo(a) and the mechanism by which protease activity was lost. As shown in Fig. 1B, the kringle domain amplified in human apo(a) is most closely related to plasminogen kringle IV, while hedgehog apo(a) kringles cluster with plasminogen kringle III. Notably, in hedgehog plasminogen it is kringle III and not kringle IV that contains the motif required for fibrin binding, supporting the independent duplication of functionally similar domains in these two species. In hedgehog apo(a) the absence of protease activity present in plasminogen is due to the deletion of the whole proteolytic domain while in primate apo(a) the proteolytic domain of plasminogen, though present, has been inactivated as a result of a mutation in the catalytic region. These observations indicate that plasminogen remodeling took place by distinct independent pathways in a subset of primates and insectivores to generate two molecules with very similar functional properties.

Sequence analysis of the plasminogen locus in mammals lacking apo(a). While the above data strongly supports convergent evolution of apo(a) in primates and insectivores, gene loss is a second possible explanation for the absence of apo(a) in other mammals. If this were the case, we would predict that the plasminogen gene duplicated before the split of hedgehog ancestors from the other placental mammals and was subsequently inactivated in all species but those presently expressing the gene. To explore this possibility, we sequenced the genomic interval containing the plasminogen locus in additional mammals with and without apo(a), such as mouse and lemur (Fig. 2). The same two genes, Slc22a3 and Map3k4, bracket

the plasminogen locus in the same 5' to 3' orientation in all species examined (human, baboon, lemur, mouse, hedgehog), with the exception that Mater and not Map3k4 brackets the right end of the locus in the hedgehog. Blast alignment of the amino acid sequence of human apo(a) to the lemur and mouse plasminogen loci did not reveal any similarity at an `expect` threshold of 10 or smaller [16] with the exception of hits to the plasminogen gene itself, indicating the complete lack of an apo(a) ortholog in these loci. In addition, querying of the complete assembly of the mouse genome also indicated the absence of apo(a) or apo(a)-like genes elsewhere in the genome of this organism. Blast alignment of the human apo(a) amino acid sequence to the mouse genome assembly version MGSCv3 returned several hits to kringle- or serine protease domain-containing regions. The best hit (e-score= 2e-27) was to a region of chromosome 9 with no annotated genes and contained one copy of a domain with 40% and 64% similarity to apo(a)'s kringle IV and proteolytic domains, respectively. The second best hit (e-score= 8e-21) was to plasminogen. None of the remaining hits (e-score= 3e-13 or greater) contained more than one copy of a kringle domain and bore no similarity to apo(a) structure. These data provide direct evidence for the absence of an apo(a) gene as the cause of the failure to detect apo(a) proteins in apo(a)-deficient species.

The baboon locus, like the human locus and in contrast to the lemur, has the apo(a) gene as well as two additional related pseudogenes, suggesting that the whole locus may have undergone several duplications at one point after the last common ancestor of lemurs and monkeys. Of note is the relative orientation of plasminogen and its neighboring gene, Slc22a3, in lemurs and baboons. It is opposite in these two primates, while apo(a) retains plasminogen's original configuration, suggesting that after the duplication the original

plasminogen gene was remodeled into apo(a) while its copy was kept as the functional plasminogen. In the hedgehog on the contrary, Slc22a3 and plasminogen are in the same configuration as in mouse and lemur. Sequence analysis of species lacking apo(a), indicating that apo(a)-like gene remnants are not present in mouse and lemur, provides further support to the absence of apo(a) not being the result of simply gene loss in these organisms.

Convergent evolution of the apo(a) gene. The species distribution of apo(a) poses an interesting evolutionary problem that can be reconciled with either gene loss in all the species lacking an apo(a) gene product or the duplicate emergence of the apo(a) gene in the distant families containing this gene. Our data, based on the sequence analysis of the genomic locus containing plasminogen and/or apo(a) in several species with (human, baboon, hedgehog) and without (mouse, lemur) and apo(a) gene product, unequivocally demonstrate that the apo(a) gene arose twice in the course of mammalian evolution. The use of different plasminogen remodeling pathways in hedgehogs and primates to produce a molecule with multiple fibrin binding sites, lacking the proteolytic activity present in plasminogen and with the ability to travel in the plasma with LDL, together with the lack of signs of remnants of plasminogen duplication in species lacking an apo(a) gene product provide very strong support to the convergent evolution of apo(a) and contradict the gene loss hypothesis. This conclusion is further supported by the partial hedgehog cDNA analysis and reconstruction of the phylogenetic tree of apo(a) in mammals [14].

Primate and insectivore apo(a) represent a remarkable example of convergent evolution, the process where molecules independently acquire a similar function [17]. A number of examples of convergent evolution are available in nature. Similar catalytic activities have

been achieved in structurally distinct enzymes through the independent evolution of the same mechanism of action in the serine proteases subtilisin and chymotrypsin [18] and in a group of methionine sulfoxide reductases [19]. Functional convergence has been described for a variety of proteins, including myoglobin in abalone and vertebrates [20] and antifreeze glycoproteins in Arctic and Antarctic fishes [21]. While all these examples have in common the evolution of similar biological functions from unrelated ancestral genes, the emergence of the apo(a) gene illustrates a different pathway for convergent evolution. In the case of apo(a), the duplication and remodeling of the same gene, plasminogen, occurred twice to separately generate paralogous genes the products of which have similar biochemical activities. Based on our lack of an understanding of the true functional demands met by apo(a)'s ability to bind to LDL and interact with fibrin, this gene further serves as a unique example of convergent evolution where the molecular footprints of a gene's arising have been revealed unaccompanied by clues to the biological need filled by its emergence.

METHODS

Construction of the hedgehog and lemur BAC libraries. The hedgehog BAC library (LBNL-4) was made from genomic DNA isolated from white blood cells of an albino African hedgehog (*Atelerix albiventris*). The lemur BAC library (LBNL-2) was constructed from genomic DNA isolated from a cell line of the ring-tailed lemur (*Lemur catta*, Coriell# AG07100C). Isolated DNA was partially digested with EcoRI and EcoRI methylase, size selected and fractionated by pulsed-field gel electrophoresis, and cloned into the pBACe3.6 vector according to the protocol described [22]. The ligated DNA was then transformed into DH10B electro-competent cells (Invitrogen). The hedgehog library contains 528 384-well plates, of which about 1% are empty or have no insert. The estimated total coverage of the hedgehog library is 10.3X, with 82% of the clones have insert ranging in size between 140 and 220kb with an average size of 156kb. The lemur library contains 291 384-well plates, of which about 5% are empty or have no insert. The estimated total coverage of the lemur library is 6.1X, with 84% of the clones have insert ranging in size between 140 and 180kb with an average size of 174kb. Both libraries are available for distribution from BacPac Resources (http://www.chori.org/bacpac/).

Library screening. The hedgehog, lemur and baboon RP41 BAC libraries were screened by filter hybridization with probes PCR-amplified from genomic DNA of the relevant species. The mouse RP23 BAC library was screened by PCR of arrayed microtiter plates. Amplification primers were either degenerate or based on the cDNA sequence of apo(a) and plasminogen. Positive clones were confirmed by PCR with additional primer pairs, and digested with BstZ17Y and size-fractionated to generate BAC contig maps to select the best candidate BAC for sequencing.

BAC sequencing and assembly. Determination of BAC sequences was carried out by the shotgun sequencing method. BAC DNA was isolated with QIAGEN Large-Construct Kit (Qiagen), randomly sheared using the Hydroshear system (GeneMachines), end-repaired and subcloned into pUC19. Fluorescence automated DNA sequencing was carried out using BygDye chemistry in an ABI3700 Sequencer (Applied Biosystems). Base calling, quality assessment and assembly was carried out using the `phred, Phrap, phrapview, Consed` software suite developed by the University of Washington Genome Center (www.phrap.org). BACs were shotgun sequenced to 8-10X coverage and subjected to one round of automated finishing with `Consed`. Contig order and orientation of non-finished assemblies was determined by a combination of paired-end analysis in `phrapview` and gene content information.

ACKNOWLEDGEMENTS

We would like to thank Sharon Massena for the generous gift of the biological specimen used for the construction of the hedgehog BAC library; Bruce Roe of the University of Oklahoma for sequencing the mouse and baboon clones under the NHGRI Trans-NIH Mouse Initiative; and Ze Peng and Keith Lewis for technical support. Research was conducted at the E.O. Lawrence Berkeley National Laboratory. The work was supported by the Grant # HL66728, Berkeley-PGA, under the Programs for Genomic Application, funded by National Heart, Lung, and Blood Institute, USA, and performed under Department of Energy Contract DE-AC0376SF00098, University of California.

REFERENCES

[1] Kraft H. G., Kochl S., Menzel H. J., Sandholzer C., and Utermann G. The apolipoprotein (a) gene: a transcribed hypervariable locus controlling plasma lipoprotein (a) concentration. *Hum Genet* **90** (1992) 220-30.

[2] Boerwinkle E., Leffert C. C., Lin J., Lackner C., Chiesa G., and Hobbs H. H. Apolipoprotein(a) gene accounts for greater than 90% of the variation in plasma lipoprotein(a) concentrations. *J Clin Invest* **90** (1992) 52-60.

[3] Utermann G., Menzel H. J., Kraft H. G., Duba H. C., Kemmler H. G., and Seitz C. Lp(a) glycoprotein phenotypes. Inheritance and relation to Lp(a)-lipoprotein concentrations in plasma. *J Clin Invest* **80** (1987) 458-65.

[4] Boffelli D., Zajchowski D. A., Yang Z., and Lawn R. M. Estrogen modulation of apolipoprotein(a) expression. Identification of a regulatory element. *J Biol Chem* **274** (1999) 15569-74.

[5] Yang Z., Boffelli D., Boonmark N., Schwartz K., and Lawn R. Apolipoprotein(a) gene enhancer resides within a LINE element. *J Biol Chem* **273** (1998) 891-7.

[6] Luc G., Bard J. M., Arveiler D., Ferrieres J., Evans A., Amouyel P., Fruchart J. C., and Ducimetiere P. Lipoprotein (a) as a predictor of coronary heart disease: the PRIME Study. *Atherosclerosis* **163** (2002) 377-84.

[7] Danesh J., Collins R., and Peto R. Lipoprotein(a) and coronary heart disease. Meta-analysis of prospective studies. *Circulation* **102** (2000) 1082-5.

[8] Sharrett A. R., Ballantyne C. M., Coady S. A., Heiss G., Sorlie P. D., Catellier D., and Patsch W. Coronary heart disease prediction from lipoprotein cholesterol levels, triglycerides, lipoprotein(a), apolipoproteins A-I and B, and HDL density subfractions: The Atherosclerosis Risk in Communities (ARIC) Study. *Circulation* **104** (2001) 1108-13.

[9] Utermann G. The mysteries of lipoprotein(a). *Science* **246** (1989) 904-10.

[10] Lou X. J., Boonmark N. W., Horrigan F. T., Degen J. L., and Lawn R. M. Fibrinogen deficiency reduces vascular accumulation of apolipoprotein(a) and development of atherosclerosis in apolipoprotein(a) transgenic mice. *Proc Natl Acad Sci U S A* **95** (1998) 12591-5.
[11] Lawn R. M., Boonmark N. W., Schwartz K., Lindahl G. E., Wade D. P., Byrne C. D., Fong K. J., Meer K., and Patthy L. The recurring evolution of lipoprotein(a). Insights from cloning of hedgehog apolipoprotein(a). *J Biol Chem* **270** (1995) 24004-9.
[12] Murphy W. J., Eizirik E., O'Brien S. J., Madsen O., Scally M., Douady C. J., Teeling E., Ryder O. A., Stanhope M. J., de Jong W. W., and Springer M. S. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* **294** (2001) 2348-51.
[13] Arnason U., Adegoke J. A., Bodin K., Born E. W., Esa Y. B., Gullberg A., Nilsson M., Short R. V., Xu X., and Janke A. Mammalian mitogenomic relationships and the root of the eutherian tree. *Proc Natl Acad Sci U S A* **99** (2002) 8151-6.
[14] Lawn R. M., Schwartz K., and Patthy L. Convergent evolution of apolipoprotein(a) in primates and hedgehog. *Proc Natl Acad Sci U S A* **94** (1997) 11992-7.
[15] McLean J. W., Tomlinson J. E., Kuang W. J., Eaton D. L., Chen E. Y., Fless G. M., Scanu A. M., and Lawn R. M. cDNA sequence of human apolipoprotein(a) is homologous to plasminogen. *Nature* **330** (1987) 132-7.
[16] Karlin S., and Altschul S. F. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A* **87** (1990) 2264-8.
[17] Cooper D. N. "Human Gene Evolution," BIOS Scientific Publishers, Oxford (1999).
[18] Kraut J., Robertus J. D., Birktoft J. J., Alden R. A., Wilcox P. E., and Powers J. C. The aromatic substrate binding site in subtilisin BPN' and its resemblance to chymotrypsin. *Cold Spring Harb Symp Quant Biol* **36** (1972) 117-23.
[19] Kryukov G. V., Kumar R. A., Koc A., Sun Z., and Gladyshev V. N. Selenoprotein R is a zinc-containing stereo-specific methionine sulfoxide reductase. *Proc Natl Acad Sci U S A* **99** (2002) 4245-50.
[20] Suzuki T., Yuasa H., and Imai K. Convergent evolution. The gene structure of Sulculus 41 kDa myoglobin is homologous with that of human indoleamine dioxygenase. *Biochim Biophys Acta* **1308** (1996) 41-8.
[21] Chen L., DeVries A. L., and Cheng C. H. Evolution of antifreeze glycoprotein gene from a trypsinogen gene in Antarctic notothenioid fish. *Proc Natl Acad Sci U S A* **94** (1997) 3811-6.
[22] Osoegawa K., Woon P. Y., Zhao B., Frengen E., Tateno M., Catanese J. J., and de Jong P. J. An improved approach for construction of bacterial artificial chromosome libraries. *Genomics* **52** (1998) 1-8.

Figure legends

Fig. 1. Panel A. Comparison of the protein structures of apo(a) and plasminogen in primates and hedgehogs. Hedgehog apo(a) domains were derived by blast alignment of hedgehog plasminogen to the genomic sequence. Kringle types are indicated in Roman numerals. Kringles with fibrin-binding ability are highlighted in purple. Hedgehog apo(a) appears to be made only of multiple copies of plasminogen kringle III. BLAST analysis of the hedgehog

apo(a) locus with the plasminogen cDNA revealed no homology to other parts of plasminogen.

Panel B. Radial tree illustrating the sequence relationships among kringle domains of human (Hs) and hedgehog (Aa) plasminogen and apo(a). Plasminogen kringle III (K3) and IV (K4) domains are highlighted in red, apo(a) domains are in blue. Only 4 copies of the multiple apo(a) kringles are displayed for human and hedgehog. Additional apo(a) kringles do not alter the structure of the tree. Phylogenetic calculations were performed using the PHYLIP package. Distances for each pair of sequences were calculated with DNADIST and used to calculate the tree using NEIGHBOR. Branch lengths are proportional to the expected number of nucleotide substitutions between nodes and are scaled so that the average rate of change over all sites analyzed is set to 1.0.

Fig. 2. Structure of the apo(a)/plasminogen locus in mammals with and without the apo(a) gene. The position and direction of the genes, indicated by the arrows, were mapped by alignment using BLAST of the respective cDNA. The tree at the right end summarizes the phylogenetic relationships among the species displayed in the figure. The sequence data are available from GenBank data library under accession nos. AC087901 (mouse), AC084862, AC084863 (baboon), AC093405 (lemur), AC122114, AC134520 (hedgehog). The human clones covering the apo(a) and plasminogen gene are under AL109933, AL596089. Contig order and orientation of the hedgehog and lemur assemblies was determined by a combination of contig paired-end analysis and electronic gene mapping to the sequence assembly.

Fig. 1A

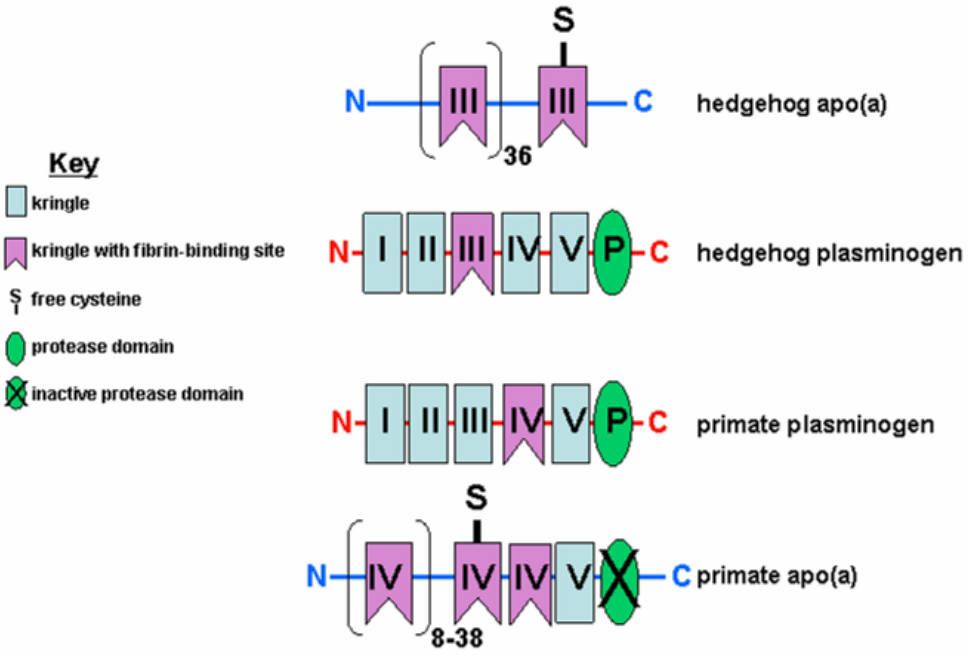

Key
kringle
kringle with fibrin-binding site
free cysteine
protease domain
inactive protease domain
S
N
III
III
C
36
hedgehog apo(a)
N I II III IV V P C
hedgehog plasminogen
N I II III IV V P C
primate plasminogen
S
N
IV
IV
IV
V
C
8-38
primate apo(a)

Fig. 1B

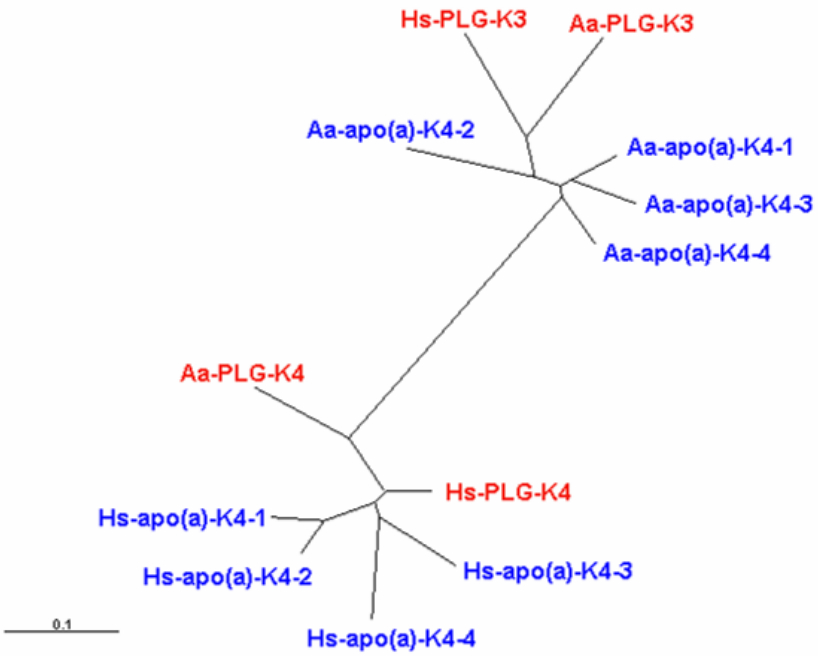
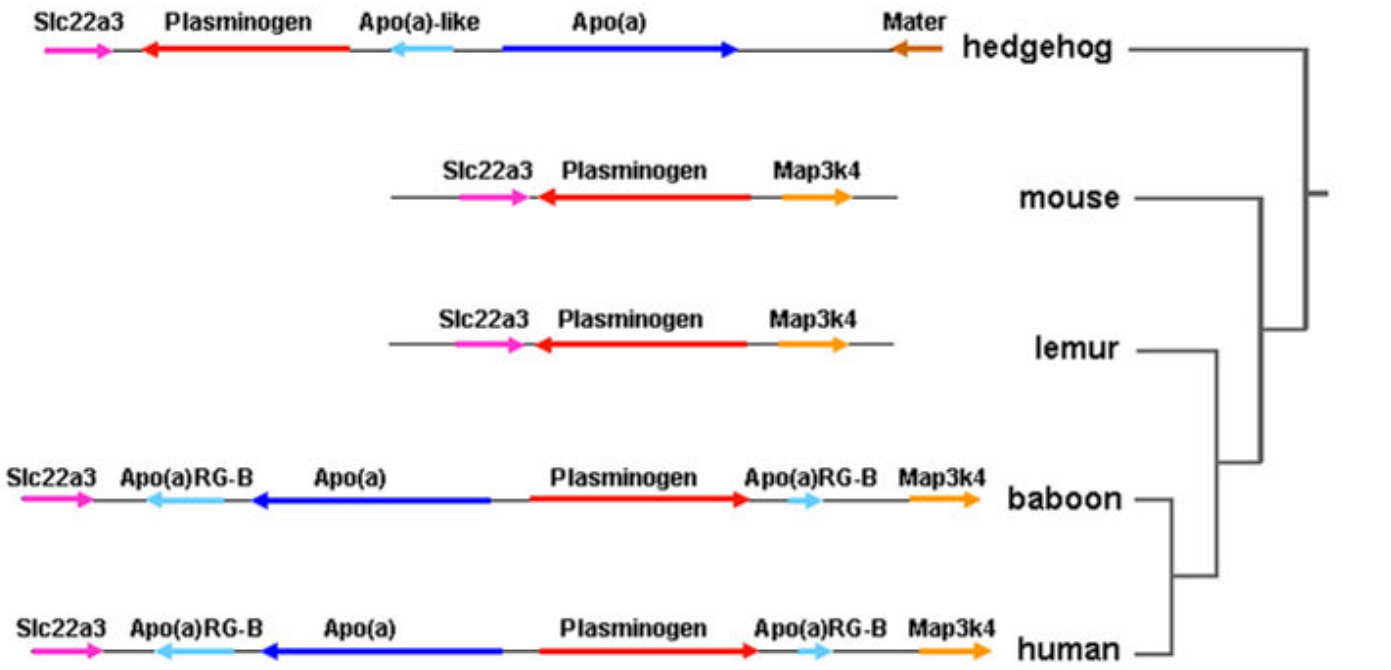

Hs-PLG-K3
Aa-PLG-K3
Aa-apo(a)-K4-2
Aa-apo(a)-K4-1
Aa-apo(a)-K4-3
Aa-apo(a)-K4-4
Aa-PLG-K4
Hs-PLG-K4
Hs-apo(a)-K4-1
Hs-apo(a)-K4-2
Hs-apo(a)-K4-3
Hs-apo(a)-K4-4
0.1

Fig. 2


Slc22a3
Plasminogen
Apo(a)-like
Apo(a)
Mater
hedgehog
Slc22a3
Plasminogen
Map3k4
mouse
Slc22a3
Plasminogen
Map3k4
lemur
Slc22a3
Apo(a)RG-B
Apo(a)
Plasminogen
Apo(a)RG-B
Map3k4
baboon
Slc22a3
Apo(a)RG-B
Apo(a)
Plasminogen
Apo(a)RG-B
Map3k4
human